BAA 07-38 FAQ's

Q: The PIP states that throughout the duration of the program, a minimum of 10,000 printed documents, 15,000 handwritten and 15,000 mixed printed and handwritten documents (the data must contain writing samples of at least 500 different writers) will need to be collected and organized. What constitutes a "document"? For some of the specified data types it seems obvious (one map, one letter) but for others it's not so clear -- for instance with newspapers, is a document one page from the newspaper; one article even if it spans multiple pages; one whole newspaper? If a map is printed on both sides, are these to be considered separate documents or a single document?

A: We will assume that a page rather than a whole newspaper would be considered an entire document. But to qualify, we would want to see a minimum number of words per document: 300 words for printed and 200 words for handwritten.

Q: Are mixed-language documents of interest, for instance a sign that contains both Arabic and French? If so, we presume that this fact should be indicated in the document metadata.

A: Yes for testing script and language ID. But the other languages will not be translated.

Q: We presume that the list of data types provided on page 4 is intended to be illustrative rather than exhaustive, and that other similar data types like pamphlets, flyers, brochures and the like would also be of interest. Could you clarify?

A: The figures are just for illustration purposes. We are interested in a variety of documents, always assuming that we get enough words to satisfy the requirement.

Q: In collecting the data for testing these systems as they are developed, will there be an even percentage of document images from photos, graffiti and signs vs. scanned-in newspapers and other traditional media?

A: We have not yet determined the specifics, but there will be a fair distribution across both printed material and handwritten material.

Q: Looking forward to the ultimate use of this system in real tactical situations where people are breaking into rooms using some device to photograph graffiti on patrols and so forth, can you explain your vision of what this scenario might look like?

A: Obviously, a solider will not be carrying a few hundred pounds of scanners on his or her back. However, most of the high level digital cameras today give you about 300 bits per image, so this will not be a problem. How a picture is taken is important; it becomes an interactive process.

Q: Will the Industry Day briefing slides be made available?

A: The slides will be posted on the Industry Day section of this website.

Q: Will there be an exception for the requirement of prepublication review for universities?
A: The DoD rule is based on the type of money used to fund the program. At this time, we don't know whether it will be 6.1 or 6.2 funding. Grants and contracts funded by 6.1, whether performed by universities or industry, or funded by 6.2 and performed on-campus at a university, are defined as "Contracted Fundamental Research (CFR)." It is DoD policy to allow the publication and public presentation of unclassified CFR results. If a university is a sub to an industry contractor under a contract funded with 6.2 money, the industry contractor is not performing CFR but its university sub is (as long as the research is on-campus). In this situation, there are no prior approvals required for the university to publish although there would be for the industrial prime.

Q: How will the metadata generation and zone labeling be evaluated, and how does the classification accuracy target apply to this? What will the targets be for metadata generation and zone labeling?

A: The classification targets are not just for type classification, but the entire set of metadata, such as logo ID and so on (as specified on slide 9 of the Industry Day slides). Basically, we will have enough documents to be able to get statistical variance so you can see how well you are doing on classification as well as translation.

Q: Are you considering animating the cat?

A: No.

Q: If a team bids on both task 1 and 2, how many cost volumes does one submit, one for task 1, one for task 2, or one for the combination of tasks 1 and 2?

A: You should submit a combined cost volume. The assumption is that the combined offer will be slightly less expensive than the separate ones.

Q: You stated that machine translation will be part of the BAA. Can you clarify its role in the evaluation as opposed to its role in GALE?

A: Machine Translation (MT) is extremely important to this program since the evaluation will only be done to the resultant English text. However, offerors should not include a budget line for MT research in their proposals. They are encouraged to use MT if they own it or team up with an MT provider. Offerors using MT (whether their own or a team-member's) should include a budget line for integration so that they can assimilate that MT with all the other processes.

Q: How much emphasis will there be on metadata vs. machine translation?

A: 80/20 or 90/10 for translation vs. metadata. Essentially what we want to be able to do is understand what is in the document. There is no task just for metadata. The tasks are for the full translation. Proposals for only parts of the process will not be seriously considered. We are looking for an end to end solution.

Q: I know that this program is not yet funded, but do you have any idea of what the budget will be and how many grants will be awarded in association with this project?

A: No. It's not a question of whether the program is funded or not. What we are interested in is extremely strong offers, and then we will know how much funding

there will be. It could be 1 or it could be 15. It absolutely depends on the strength of the offers.

Q: Is evaluation going to be based on single reference translation?

A: It will be based on a gold standard, which is a single transcript, but it is a composite. We use 2 or 3 translators to start with; then conduct quality assurance; then an adjudicator looks at the whole ensemble and writes a composite translation. The adjudicator decides if there is ambiguity and if so, it will be indicated in the gold standard. It is as close to perfect translation as one can get.

Q: Are errors of fluency and word choice important? Will there be penalties for picking a synonym?

A: Fluency is obviously important. Lack of fluency that interferes with the meaning being conveyed will not be acceptable. As long as synonyms are

correct, they will be fine. We use humans in the process of deciding which synonyms are acceptable. The editors, who are extremely well trained, compare the output to the gold standard. They assess how well the meaning is conveyed. This is as fair of a process as we can come up with.

Q: Will the data types be increasingly challenging for each of the phase or is the phase I corpora representative of the rest of the phases?

A: The real challenge is constant improvement. To measure improvements, the test corpora will be of equivalent difficulty.

Q: When will the phase I data become available?

A: We will have to determine that with the offers.

Q: Will this only include hardcopy materials like paper or will there be web data such as website translation as well?

A: In MADCAT, we will not be dealing with Unicode text input at all. MADCAT will not include the genres like email that are in the conversational mode.

Q: Is there a task for quality control? Are you going to develop a quality control testing scenario, or do you have one already?

A: Within the program, there is the determined metric that was stated in the BAA. This metric has been approved by the DARPA director. There are shortcomings in this method since it weighs all words equally. However, when we get to our ultimate target of 95% accuracy, this won't matter. We are asking the machines to do better than your normal level 2 or 3 translator and we expect to succeed. As far as the gold standard is concerned, it goes through a tremendous amount of scrutiny.

Q: Are you looking to develop a metric system to measure the success of this or do you already have one?

A: We are not looking for offers to develop new testing paradigms. We are looking for someone to do the evaluation by the methods that we have described.

Q: Are you expecting bids to propose a 5 phase program, and should we assume that the length of each phase will be 1 year?

A: I am expecting the offerors to show methods to get to the ultimate goal, and it is assumed that proposals will be for 5 phases. You are welcome to offer a shorter time period as well as shortened phase periods. 12 months is dictated by the DARPA director as the highest limit for the duration of each phase. The intermediate targets are for you to offer in the proposals, keeping in mind that each phase's targets must be met in order to be eligible to proceed to the next phase. However, DARPA may eliminate some performers before the end of the program even if they have met the targets for a given phase.

Q: Have you picked a specific Arabic dialect, or are we looking at the entire range of dialects?

A: Printed material will be mostly in MSA. Levantine and Iraqi will be the most likely ones that MADCAT will include. We will not be working with North African dialects, which are quite far from Middle East dialects.

Q: On page 8 of the BAA, it speaks about sharing of the technical details, especially with respect to algorithm development. If we were to build on a commercial product, does that requirement include the intellectual property of the initial product?

A: We would allow anyone to file for a provisional patent before disclosing. We will respect all intellectual property rights.

Q: Would DARPA be interested in a classified evaluation component?

A: If we encounter classified material, we will have to classify parts of the program. We would prefer, however, to keep it open.

Q: There is a section on export licenses that says that we must control access by foreign persons to export controlled technical data and software. Will this prevent students from certain countries from working at universities on research funded under this program?

A: No, this only prevents the foreign student from working on the project without first having the university obtain an export license for that country for the specific technology. This is not a show stopper - but the university must apply for and receive the export license BEFORE the student is allowed to participate in the project.

Q: Will one of the surprise languages be one of the top 20 resource-rich languages?

A: Either that or another language for which we could provide you enough data to train your algorithms on.

Q: Will there be other types of documents than printed documents and handwritten paper documents like graffiti on walls? And if so, would they be part of the training data?

A: There will not be anything evaluated for which there was not sufficient training data. The details will be determined when we can assess the availability of data.